From Humans to Machines: Can ChatGPT-like LLMs Effectively Replace Human Annotators in NLP Tasks?

Surendrabikram Thapa[1], **Usman Naseem**[2], **Mehwish Nasim**[3,4]

[1] Department of Computer Science, Virginia Tech, USA
[2] School of Computer Science, The University of Sydney, Australia
[3] School of Physics, Mathematics and Computing, The University of Western Australia, Australia
[4] College of Science and Engineering, Flinders University, Australia
surendrabikram@vt.edu, usman.naseem@sydney.edu.au, mehwish.nasim@uwa.edu.au

Abstract

The increasing demand for natural language processing (NLP) applications has created a need for large amounts of labeled data to train machine learning models. This has led to using human annotators for tasks such as text classification, sentiment analysis, and named entity recognition. However, human annotation is costly and time-consuming, and the annotation quality can significantly vary depending on the annotator. Recent advances in language modeling have led to the development of large language models (LLMs), such as ChatGPT, which are capable of generating human-like responses to text prompts. In this position paper, we explore the question: *whether ChatGPT-like LLMs can effectively replace human annotators in NLP tasks?* We discuss the advantages and limitations of using LLMs for annotation and highlight some of the challenges that need to be addressed to make this a feasible approach. We argue that while LLMs can potentially reduce the cost and time required for annotation, they may not be able to fully replace human annotators in all NLP tasks. We conclude by outlining future research directions that could help advance the use of LLMs for NLP annotation.

Background

The recent emergence of large language models (LLMs) such as GPT-3.5, GPT-4, and LLaMA has been a significant catalyst for the progress of natural language processing (NLP). These LLMs have been instrumental in achieving exceptional results across various NLP tasks, highlighting their immense potential in this field. Despite their exceptional performance, the practical application of LLMs in production environments is often limited by factors such as their size, slow inference speed, and high cost. Additionally, the lack of publicly available parameters for such large language models can restrict their flexibility for local deployment. As a result, smaller language models like BERT are often preferred for use in production settings where efficiency and cost-effectiveness are important considerations. These smaller models can still achieve high performance on various NLP tasks while being more efficient and cost-effective than larger models. Kocoń et al. (2023) suggested that LLMs like ChatGPT are lagging behind the currently available SOTA models by 4 to 70% when tested on 25 different NLP tasks. This clearly highlights the need to improve SOTA models which are designed to be task-specific in order to achieve better performance on various NLP tasks. Nevertheless, it is important to note that the performance of these task-specific models heavily depends on the quality of the annotated data used for their training and fine-tuning. However, data annotation can be a costly undertaking, as it is costly and time-consuming. Given the challenges associated with data annotation, there has been growing interest in exploring the potential of using large language models like ChatGPT to automate the annotation process. The workflow of the NLP research can be optimized drastically if data annotation can be automated.

While LLMs have demonstrated exceptional performance in annotation for various NLP tasks, they are not without their limitations. For instance, these models can be biased, and their predictions can be influenced by the training data they were exposed to. Additionally, these models may struggle to accurately identify nuances and context-specific meanings in language, leading to errors and incorrect annotations. Therefore, it is not yet clear whether these models can effectively replace human annotators in NLP tasks. This position paper explores the current practices, challenges, and potential future directions in using large language models like ChatGPT for automated data annotation.

Annotation Practices with LLMs

In recent studies, researchers have evaluated the performance of GPT-3 as an annotator in various NLP tasks. According to Ding et al. (2022), GPT-3 is well-suited for text classification tasks due to its large-scale pre-training, making it ideal for directly tagging test data. However, for more complex tasks such as Named Entity Recognition (NER), prompt-guided or dictionary-assisted training data generation may be more effective and cost-efficient. Another study by Wang et al. (2021) investigated the efficiency of GPT-3 annotations in various NLU and NLG tasks. While the study by Brown et al. (2020) proposed the direct use of GPT-3 for downstream tasks, the findings from Wang et al. (2021) suggest that in-house models such as PEGASUS$_{large}$ and RoBERTa$_{large}$ perform better when trained on datasets labeled with GPT-3 annotations. This implies that relying solely on GPT-3 for downstream tasks may not yield the best results, and using it to generate labeled datasets may be more

effective in improving the performance of other models. Additionally, using LLMs like GPT-3 directly can be costly and may have higher latency in real-world applications.

With the promising abilities of ChatGPT, researchers have been working to explore the possibilities of annotating with ChatGPT. Recent studies such as Mei et al. (2023) have shown promising results in using ChatGPT as an assistive tool for annotation tasks, while others such as Gilardi, Alizadeh, and Kubli (2023) and Huang, Kwak, and An (2023) have explored the possibility of ChatGPT performing annotations on its own, demonstrating the ChatGPT's potential to improve annotation efficiency and accuracy. Huang, Kwak, and An (2023) evaluated the efficacy of ChatGPT in the annotation of 795 implicit hateful tweets. The responses of ChatGPT not only included the labels but also the explanations (NLEs) which were assessed in the study. ChatGPT was able to recognize 80% of the tweets as hateful. When the NLEs were assessed for informativeness and clarity, NLEs generated by ChatGPT outperformed human beings significantly for clarity scores and had similar informativeness scores. When the original annotators were given the tweets with ChatGPT-based NLEs, the hatefulness score was influenced significantly. The original annotators changed some of the labels to non-hateful which means that the people's decisions can be influenced by additional explanations given by ChatGPT. Similarly, in another study by Gilardi, Alizadeh, and Kubli (2023), ChatGPT outperformed crowd-workers when a comparison was made with a sample of 2,382 tweets.

Challenges

Using large language models (LLMs) like ChatGPT for data annotation can present several challenges for the computational social science community. These models have the potential to automate the annotation process, but several factors need to be considered to ensure the accuracy and reliability of the annotations. LLMs require large amounts of high-quality training data to learn effectively. The quality of the training data can affect the accuracy and reliability of the annotations generated by the model. The models can inherit biases present in the training data, which can result in biased annotations. Additionally, the models may not be fair and could result in unequal treatment of certain groups. LLMs are often considered black boxes, meaning it can be difficult to understand how the model is making its annotations. This lack of transparency can make it difficult to identify errors or biases in the annotations. Human annotators are capable of controlling bias in annotation, whereas LLMs lack this ability. LLMs trained on general language tasks may not have the domain-specific knowledge required for certain annotation tasks, which could lead to inaccurate or irrelevant annotations. The annotations for sensitive data, for example, medical data are difficult for general language models. With the advent of domain-specific language models like BloombergGPT (Wu et al. 2023), annotation abilities for various specific domains are yet to be explored. LLMs may struggle with complex linguistic constructions, such as sarcasm, irony, and metaphor, which could affect the accuracy of the annotations. The use of LLMs for data annotation may raise concerns about the ownership and privacy of the annotated data. There may also be legal considerations related to the use of personal data for annotation, particularly in cases where the data contains sensitive information.

Direction Ahead

Recent studies have shown that large language models (LLMs) like ChatGPT can assist in data annotation and may even be able to replace human annotators to some extent. However, it is crucial to maintain human-in-the-loop validation during the annotation process to ensure the reliability and accuracy of the annotations. One approach that has been proposed is active labeling, where humans re-label low-confidence instances provided by the LLMs (Wang et al. 2021). Another direction is to develop LLMs specifically for annotation tasks, with domain-specific knowledge and specialized training data. This could improve the accuracy and relevance of annotations for specific tasks. Additionally, LLMs could also be leveraged to identify and correct errors or biases in existing annotated data, thereby improving the quality of datasets. In summary, while LLMs like ChatGPT have the potential to automate the annotation process, it is important to consider the challenges and limitations to ensure the reliability and accuracy of the annotations.

References

Brown, T.; Mann, B.; Ryder, N.; Subbiah, M.; Kaplan, J. D.; Dhariwal, P.; Neelakantan, A.; Shyam, P.; Sastry, G.; Askell, A.; et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33: 1877–1901.

Ding, B.; Qin, C.; Liu, L.; Bing, L.; Joty, S.; and Li, B. 2022. Is GPT-3 a Good Data Annotator? *arXiv preprint arXiv:2212.10450*.

Gilardi, F.; Alizadeh, M.; and Kubli, M. 2023. ChatGPT Outperforms Crowd-Workers for Text-Annotation Tasks. *arXiv preprint arXiv:2303.15056*.

Huang, F.; Kwak, H.; and An, J. 2023. Is ChatGPT better than Human Annotators? Potential and Limitations of ChatGPT in Explaining Implicit Hate Speech. *arXiv preprint arXiv:2302.07736*.

Kocoń, J.; Cichecki, I.; Kaszyca, O.; Kochanek, M.; Szydło, D.; Baran, J.; Bielaniewicz, J.; Gruza, M.; Janz, A.; Kanclerz, K.; et al. 2023. ChatGPT: Jack of all trades, master of none. *arXiv preprint arXiv:2302.10724*.

Mei, X.; Meng, C.; Liu, H.; Kong, Q.; Ko, T.; Zhao, C.; Plumbley, M. D.; Zou, Y.; and Wang, W. 2023. WavCaps: A ChatGPT-Assisted Weakly-Labelled Audio Captioning Dataset for Audio-Language Multimodal Research. arXiv:2303.17395.

Wang, S.; Liu, Y.; Xu, Y.; Zhu, C.; and Zeng, M. 2021. Want to reduce labeling cost? GPT-3 can help. *arXiv preprint arXiv:2108.13487*.

Wu, S.; Irsoy, O.; Lu, S.; Dabravolski, V.; Dredze, M.; Gehrmann, S.; Kambadur, P.; Rosenberg, D.; and Mann, G. 2023. BloombergGPT: A Large Language Model for Finance. arXiv:2303.17564.